HIT Standards Committee – Clinical Operations Workgroup – Task Force on Vocabulary
Panel 3: Best Practices & Lessons Learned: Vocabulary Infrastructure
March 23, 2010
Testimony by Stuart Nelson, M.D., National Library of Medicine, National Institutes of Health

Thank you for the opportunity to present to you on behalf of the National Library of Medicine (NLM). I am Dr. Stuart Nelson, Head of the Medical Subject Headings Division at NLM. In addition to being in charge of the maintenance and oversight of MeSH®, our vocabulary for indexing the medical literature, I am in charge of maintenance of the Unified Medical Language System® (UMLS®) Metathesaurus®, direct the development of RxNorm®, and oversee NLM's contribution to SNOMED CT® content development.

Background

The National Library of Medicine has been in the electronic vocabulary creation and distribution business for about 50 years, from the start of MeSH, and has been involved in the dissemination of clinical vocabularies and administrative code sets for more than 20 years via the UMLS. The UMLS project strongly influenced terminology research and the definition of what constitutes a well-formed vocabulary. At some level, the UMLS resources underpin virtually all significant natural language processing efforts involving English language biomedical or clinical free text.

NLM is the US Member of the International Health Terminology Standards Development Organisation (IHTSDO), which owns SNOMED CT. The Library supports development, ongoing maintenance and free US-wide access to SNOMED CT, RxNorm, and LOINC®. NLM also produces the NCBI Taxonomy and RefSeqGene**,** a reference standard for identifying the location of clinically-significant genetic variation.

1) **What vocabulary subset or value set creation and distribution services do you provide?**
2) **Who uses your services and what is the level of use?**
3) **What, if any, additional services and capabilities are in active development?**
4) **What process is used to establish and revise the subsets or value sets that you distribute?**

NLM provides and supports three general categories of services:

a) tools that allow a wide range of users (researchers, information service developers, and system developers, among others) to identify and obtain in a uniform fully-specified format appropriate vocabularies or parts of vocabularies in many different languages;
b) creation and distribution of subsets of standard clinical vocabularies that cover frequently used clinical concepts as a convenience for those implementing electronic health records;
c) attachment of standard codes and terminology to value sets created or identified by other groups.

<u>UMLS Metathesaurus</u>

The Unified Medical Language System (UMLS), with its Knowledge Sources, is the primary means by which NLM distributes vocabularies, vocabulary subsets and value sets. The UMLS

Metathesaurus is a very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them. The UMLS Metathesaurus contains almost 150 source vocabularies and provides a value-added dissemination mechanism for key terminology, classification, and coding standards. It distributes them in a common fully-specified format with consistent semantic categorization, establishing synonymous relationships among them, and facilitating their use in conjunction with UMLS lexical tools. The Metathesaurus is updated twice a year. Its constituent vocabularies are updated by their developers on their own schedules.

The Metathesaurus reflects and preserves the meanings, concept names, and relationships from its source vocabularies. In other words, the Metathesaurus does not represent a comprehensive NLM-authored ontology of biomedicine or a single consistent view of the world. Using "concepts" to link across terminologies, the Metathesaurus preserves the many views of the world present in its source vocabularies. These different views are useful for different tasks.

The large size and scope of the UMLS Metathesaurus necessitated the development of tools for creating user-defined subsets. The first tool, the Content View Flag (CVF), is a mechanism that enables users to mark and extract subsets. The Content View mechanism is used to pre-compute subsets of the Metathesaurus (and its component vocabularies) which are then instantiated during production. A companion resource, the UMLS installation and customization tool MetamorphoSys, may be used to exclude vocabularies that are not required or licensed for use in local applications and to select from a variety of data output options and filters, e.g., by language, semantic type or general category. MetamorphoSys can also be used to select out a particular CVF for use. Users can also access the UMLS Metathesaurus via the UMLS Knowledge Source Server, which has both browser and Applications Program Interfaces.

The UMLS Metathesaurus is available free under the terms of a license that specifies restrictions that apply to some of its source vocabularies. There are 4,700 licensees worldwide.

Clinically Relevant Subsets

NLM has developed or contributed to the development of clinically relevant vocabulary subsets for problems, medications, and clinical observations.

Problems

The CORE Problem List Subset of SNOMED CT is an output of the UMLS CORE Project (CORE stands for **C**linical **O**bservations **R**ecording and **E**ncoding), which intends to define a UMLS subset that would be most useful for documentation and encoding of clinical information at a summary level, such as problem list, discharge diagnosis or reason for encounter. The approach is to collect and analyze datasets from health care institutions that currently use controlled vocabularies for data entry. These datasets contain the list of controlled terms and their actual frequency of usage in the health care institutions.

The CORE Problem List Subset was derived based on datasets from 7 institutions: Beth Israel Deaconess Medical Center, Intermountain Healthcare, Kaiser Permanente, Mayo Clinic, Nebraska University Medical Center, Regenstrief Institute and Hong Kong Hospital Authority. These institutions are large-scale, mixed inpatient-outpatient facilities that cover most major medical specialties (including Internal Medicine, General Surgery, Pediatrics, Obstetrics, Gynecology, Psychiatry and Orthopedics). The most frequently used terms, about 14,000 in all, represented about 95% of the usage volume in each institution. These were mapped to 6,800 UMLS concepts, which formed the basis of the UMLS CORE subset. (Only about 8% of terms were not mappable to the April 2009 edition of the UMLS Metathesaurus.) We were pleased but not surprised to see that, among the source terminologies in the UMLS, SNOMED CT covered the highest percentage (81%) of the identified UMLS CORE concepts. Relevant concepts not yet in SNOMED CT are being submitted for addition.

The main purpose of the CORE SNOMED CT Subset is to facilitate the use of SNOMED CT as the primary coding terminology for problem lists or other summary level clinical documentation. NLM believes that pairing the central clinical interaction – the statement of the patient problem -- with use of a common list of SNOMED CT concepts will ultimately maximize data interoperability among institutions. The CORE Subset is an intramural NLM product that we expect to update four times per year, in conjunction with and reflecting each new release of SNOMED CT and the UMLS. Concepts will be added or retired based on changes in SNOMED CT, evaluation and user feedback and additional analysis of source datasets.

We believe that the CORE Subset covers a high proportion (90% or more) of usage volume in general practice institutions, but users may need to expand or modify it to cover specialized, less commonly-encountered problems specific to their needs.

NLM is currently evaluating the new IHTSDO Workbench (used for editing and maintaining SNOMED CT content) for internal use, and considering ways of enhancing its core functionality with access to UMLS information. One pilot project would maintain and export the CORE Subset as a SNOMED refset using the Workbench.

We do not currently have any usage statistics for the CORE subset, which was introduced last July.

Medications

RxNorm provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management and drug interaction software, including those of First Databank, Micromedex, MediSpan, Gold Standard Alchemy, and Multum. By providing links between these vocabularies, RxNorm can mediate messages between systems not using the same software and vocabulary. NLM issues weekly updates and full monthly releases of RxNorm.

NLM is actively developing an e-prescribing subset, which would include only drugs for humans that are currently on the market in the US. This subset would be considerably smaller than the totality of RxNorm, which includes some foreign drugs, some off the market drugs, and veterinary drugs (also recently added to the Structured Product Label distribution via the DailyMed.)

NLM has signed a Memorandum of Understanding with the Veterans Administration will allow the inclusion of NDF-RT in the RxNorm distribution beginning in June 2010. Integration of RxNorm with NDF-RT will link drugs to therapeutic classes, indications, and pharmacokinetic properties. Some drug-drug interaction information will be included as well. We believe this will continue to enhance the usefulness of RxNorm, which already includes NDC codes and UNII codes from the FDA.

The NDC codes list incorporated in RxNorm is at least as complete and accurate as any available from other public or private sources. NLM has about 500 users who download RxNorm at least once a month. CMS is using RxNorm names and codes to evaluate formularies for Medicare Part D prescribing.

RxNorm can be accessed directly or through RxTerms, an NLM-developed interface terminology derived from RxNorm, intended for use in U.S. systems for prescription writing or medication history recording. RxTerms can be considered a subset of RxNorm: tailored for U.S. prescribing, RxTerms retains the broad coverage of RxNorm for U.S. prescribable drugs – 99% coverage of both generic and brand names of U.S. most commonly prescribed drugs. As of December 2009, RxTerms had about 300 registered users, including software vendors, researchers, and health care providers. Regenstrief Institute and Siemens Corporation have current developmental efforts to use RxTerms for new order entry systems. CMS is using RxTerms in a demonstration project of its post-acute care assessment tool (CARE). NLM uses RxTerms in the NLM Personal Health Record.

There is an RxNorm browser, called RxNav, which also has two Application Programming interfaces (APIs) and links to external sources of drug information, such as the DailyMed.

Usage of RxNav has increased steadily over time, reaching an monthly average of over 50,000 queries in 5000 sessions (browser and APIs). The API is used in applications including MyMedicationList and MyRxPad (e-prescribing). Based on feedback from users, RxNav and the APIs have been used in academic environments (including CTSAs), in health insurance companies, by EHR vendors, and drug information providers. Mapping NDC codes to RxNorm concepts is one of the main uses of the API, which has been employed to process large amounts of queries.

Clinical Observations

The purpose of LOINC® is to facilitate the exchange and pooling of clinical results for clinical care, outcomes management, and research by providing a set of universal codes and names to identify laboratory and other clinical observations. The Regenstrief Institute, Inc, an internationally renowned healthcare and informatics research organization, owns and maintains the LOINC database and supporting documentation, and the RELMA mapping program. NLM has provided partial support for the development and free dissemination of LOINC for more than 10 years. The Regenstrief Institute, Inc. and the IHTSDO are in the final stages of negotiation of an agreement intended to specify prospective distribution of labor in the development of LOINC and SNOMED CT and to enable combined distributions of LOINC and SNOMED CT content.

LOINC is used worldwide by local hospitals and laboratories, regional health networks (city, state, and national levels), public health departments, health care provider networks (e.g. Partners Healthcare and Intermountain Health Care), software vendors, payers, and insurance companies. Last year, there was an average of 1000 downloads of the database (either alone or with RELMA) per month or 1200 per year from a total of 136 different countries. The Regenstrief

Institute has established an online directory on their web site that allows LOINC adopters to publish their contact information and use cases. At the beginning of 2009, there were 64 organizations, institutions, and other LOINC adopters listed on this site.

NLM has worked with the Regenstrief Institute, Inc. and other interested parties on the following subsets:

- **Universal common laboratory order LOINC codes** - The purpose of this subset is to define a set of universal laboratory order codes that would be recommended and adopted by developers of order-entry systems for delivery in HL7 messages to laboratories, where they could be understood and fulfilled. The subset is designed to cover greater than 95% of the test ordering volume in the US. It was developed with both empirical and consensus-driven approaches.

- **Universal common laboratory results/observation LOINC codes** – We currently have a list of the 2000 most frequently reported laboratory *results* and the frequency-based observation data from about 350 million test results obtained from United Health Care, Partners and INPC (Indiana). The data represent 99.8% of the test volume in those systems. Prioritizing the mapping effort on the modest subset of most common laboratory result codes is a convincing way to lower the barriers to interoperability in a way that reduces implementation effort without dramatically sacrificing the rewards.

- **Database of raw unit strings mapped to UCUM** (Unified Code for Units of Measure) – UCUM is a terminology for encoding units of measure, creating a computable code for unit strings. It is recommended by most standard organizations and in the HL7 laboratory message standard. The purpose is to facilitate unambiguous electronic communication of quantities together with their units for easy inter-conversion of units. Based on the information provided by 23 sources (115,000 records) that mapped their test offerings to LOINC, we have obtained the raw unit strings reported in HL7 messages and mapped them to UCUM units. This mapping will facilitate the use of UCUM units in laboratory reporting.

Promoting the Use of Standard Codes and Terminology in Value Sets

NLM has worked with a variety of organizations and groups to ensure that key value sets are expressed in – or connected to - standard codes and terminology.

- **Newborn Screening Coding and Terminology Guide** – This guide conveys both the set of standard codes and terminology for newborn screening test measurements and the conditions for which newborns are screened in all US newborn screening programs. It includes the LOINC terms required to report all newborn screening results for all U.S. newborn screening programs, including variables for reporting an overall summary, for most of the card variables and, for reporting impressions, narrative guidance and measures of quantitative markers for each condition or condition category. The answer lists that report conditions are also coded in SNOMED CT. The LOINC NBS panel specifies the codes for an NBS HL7 message. To demonstrate how these codes load into such a message, we created an annotated example HL7 v2.5.1 NBS message, which can be found on the NLM website. This approach has been adopted by the three main commercial NBS IT system vendors who have already shown us compliant (or near compliant) prototype HL7 messages.

- **Post-acute Care Assessment Instruments --** We created a LOINC panel (template) for the Centers for Medicare and Medicaid Services (CMS) post-acute care assessment instrument. Improving the capability to use electronically-transmitted data to monitor vulnerable patients as they pass between and among hospital, home health and skilled nursing settings has long been a high priority for CMS**.** The templates have been completed and mapped to LOINC, and are available in the current release of LOINC and as a separate spreadsheet downloadable from Regenstrief Institute.

- **Clinical Genomics** – A fully LOINC-qualified HL7 Version 2 Implementation Guide and Genetic Variation Model was recently released that uses LOINC code for observations and SNOMED CT codes for diagnoses. A similar HL7 Implementation Guide, on Cytogenetics, is currently under development.

- **Integration of genetics and epidemiologic research –** NLM has been working with PhenX, which is a three-year project led by RTI International and funded by the National Human Genome Research Institute (NHGRI) to contribute to the integration of genetics and epidemiologic research. Our purpose is to ensure that data from these clinical research activities, which will collected and shared electronically as part of the NIH's commitment to increasing access to genomic data, will be encoded using standardized terminologies, such as LOINC and SNOMED CT.

5) **Based on your experience, what advice would you offer regarding best practices and pitfalls to avoid?**

Best practices and pitfalls for subsets and value sets are very similar to those that apply to standard vocabularies and code sets as a whole. Our experience leads us to the following observations:

- Standard vocabularies, generally useful subsets, and standard value sets must be supported for the long-term. Sustained committed investment in infrastructure and maintenance is critical.

- The development, maintenance, and versioning of subsets and value sets should be tightly coupled with the ongoing maintenance of the standard vocabularies and code sets from which they are drawn. Generating new subsets or value sets from outdated versions of terminologies is counterproductive. Changes in base vocabularies and code sets will necessitate updates to subsets and value sets. The development of automated methods for detecting when changes in base vocabularies and code sets are likely to affect particular subsets and value sets is a worthy topic for research and development.

- It is very important to maintain version information for a subset or value set, which in turn must maintain version information for the terminologies on which it is based. Any terminology service that is used to query for value set membership must also be version aware.

- There should be standard methods and processes for performing quality assurance on a new version of any standard value set or subset before it is published as the current version. (Our experience with terminologies has demonstrated that there are often multiple problems with a released version of a terminology. Some of these may be relatively trivial, such as an unusual character or additional white space; others are more serious, such as a violation of relational integrity.)

- Updates to standard terminologies and code sets also may trigger the necessity to update or reprocess information in electronic health records and clinical data warehouses

- The optimum tools and platform for *producing and maintaining* particular subsets and value sets may vary, in part due to the desirability of coupling these activities to the maintenance of the underlying terminologies and code sets and to differing governance mechanisms.

- A "create once" and "publish many times in many places and formats" approach to the dissemination of terminologies, code sets, and value sets may meet more needs of more stakeholders than any single distribution point and format could. Nevertheless, at a minimum there should be a central place to find out about all the vocabulary content that is relevant to meaningful use and at least one uniform way to get access to all of it.

- At present, key terminologies have incompatible, inconsistent update schedules, some of which are defined in legislation. Modification and synchronization of update schedules of key terminologies may be essential to efficient operation of any public repository of subsets and value sets that use them. Coordination will not be a simple task, since some of these changes will require Congressional action (notably the update schedules for ICD and CPT).

- Any effort to develop a repository for terminologies, value sets, and subsets should not start from scratch. Considerable investments have been made by a number of Federal agencies in order to fulfill their varying mission requirements for vocabulary dissemination. The lessons learned over time do not need to be repeated.